New normality test in high dimension with kernel methods

Jérémie Kellner, Alain Celisse

New normality test in high-dimension with kernel methods

Jérémie Kellner
Laboratoire de Mathématiques UMR 8524 CNRS - Université Lille 1 - MODAL team-project Inria
jeremie.kellner@ed.univ-lille1.fr
Alain Celisse
Laboratoire de Mathématiques UMR 8524 CNRS - Université Lille 1 - MODAL team-project Inria
celisse@math.univ-lille1.fr

Abstract

A new goodness-of-fit test for normality in high-dimension (and Reproducing Kernel Hilbert Space) is proposed. It shares common ideas with the Maximum Mean Discrepancy (MMD) it outperforms both in terms of computation time and applicability to a wider range of data. Theoretical results are derived for the Type-I and Type-II errors. They guarantee the control of Type-I error at prescribed level and an exponentially fast decrease of the Type-II error. Synthetic and real data also illustrate the practical improvement allowed by our test compared with other leading approaches in high-dimensional settings.

1 Introduction

Dealing with non-vectorial data such as DNA sequences often requires defining a kernel [1]. Further analysis is then carried out in the associated Reproducing Kernel Hilbert Space (RKHS) where data are often assumed to have a Gaussian distribution. For instance supervised and unsupervised classification are performed in [4] by modeling each class as a Gaussian process. This key Gaussian assumption is often made implicitly as in Kernel Principal Component Analysis [20] to control the reconstruction error [13], or in [18] where a mean equality test is used in high-dimensional setting. Assessing that crucial assumption appears necessary.

Depending on the (finite or infinite dimensional) structure of the RKHS, Cramer-von Mises-type normality tests [12, 9, 19] can be applied. However these tests become less powerful as dimension increases (see Table 3 in [19]). An alternative approach consists in randomly projecting high-dimensional objects on one-dimensional directions and then applying univariate test on a few randomly chosen marginals [5]. However such approaches also suffer a lack of power (see Section 4.2 in [5]). More specifically in the RKHS setting,

[6] introduced the Maximum Mean Discrepancy (MMD) and design a statistical test to distinguish between the distribution of two samples. However this approach requires *characteristic kernels* [7] and suffers high computational complexity as well as several approximations of the asymptotic distribution.

The main contribution of the present paper is to provide an algorithmically efficient one-sample statistical test of normality for data in a RKHS (of possibly infinite dimension). However the strategy we describe can be easily extended to the two-sample setting. Section 2 introduces goodness-of-fit tests available in high dimensional settings. They will serve as references in our simulation experiments. The new goodness-of-fit test is described in Section 3, while its theoretical performance is detailed in Section 4 in terms of control of Type-I and Type-II errors. Finally results of experiments on synthetic and real data highlight the great theoretical and practical improvement allowed by the new statistical test. They are collected in Section 5.

2 High-dimensional goodness-of-fit tests

2.1 Statistical test framework

Let $(\mathcal{H}, \mathcal{A})$ be a measurable space, and $Y_1, \ldots, Y_n \in \mathcal{H}$ denote a sample of *independent and identically distributed* (*i.i.d.*) random variables drawn from an unknown distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is a set of distributions defined on $\mathcal{A}$.

Following [11], let us define the null hypothesis $H_0 : \ P \in \mathcal{P}_0$, and the alternative hypothesis $H_1 : \ P \notin \mathcal{P} \setminus \mathcal{P}_0$ for any subset $\mathcal{P}_0$ of $\mathcal{P}$. The purpose of a statistical test $\mathcal{T}(Y_1, \ldots, Y_n)$ of H_0 against H_1 is to distinguish between the null (H_0) and the alternative (H_1) hypotheses. For instance if $\mathcal{P}_0$ reduces to a univariate Gaussian distribution with mean μ_0 and variance σ_0^2, $\mathcal{T}(Y_1, \ldots, Y_n)$ determines whether $P = \mathcal{N}(\mu_0, \sigma_0^2)$ is true or not for a prescribed level of confidence $0 < \alpha < 1$.

2.2 Projection-based statistical tests

In the high-dimensional setting, several approaches share a common projection idea dating back to the Cramer-Wold theorem extended to infinite dimensional Hilbert space.

Proposition 2.1. *(Prop. 2.1 from [5]) Let $\mathcal{H}$ be a separable Hilbert space with inner product $< \cdot, \cdot >$, and $Y, Z \in \mathcal{H}$ denote two random variables with respective Borel probability measures P_Y and P_Z. If for every $h \in \mathcal{H}$*

$$< Y, h > = < Z, h > \ \textit{in distribution} \ ,$$

then $P_Y = P_Z$.

Since considering all possible directions h is impossible with high-dimensional $\mathcal{H}$, [5] suggest to randomly choose some of them from a Gaussian measure. Given an *i.i.d.* sample

$Y_1, \ldots, Y_n$, a Kolmogorov-Smirnov test is performed from $< Y_1, h >, \ldots, < Y_n, h >$ for each h, leading to the test statistic

$$D_n(h) = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)| \ ,$$

where $\hat{F}_n(x)$ is the empirical cdf of $(< Y_i, h >)_i$ and F_0 denotes the cdf of the $< Z, h >$.

Since [5] proved too few directions lead to a less powerful test, this can be repeated for several randomly chosen directions h, keeping then the largest value for $D_n(h)$. However the test statistic is no longer distribution-free (unlike the univariate Kolmogorov-Smirnov one) when the number of directions is larger than 2. Therefore for a given confidence level on the Type-I error, the test threshold (quantile) must be estimated through Monte-Carlo simulations.

2.3 The Maximum Mean Discrepancy (MMD)

Following [6] the gap between two distributions P and P_0 can be measured by

$$\Delta(P, P_0) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{Y \sim P} f(Y) - \mathbb{E}_{Z \sim P_0} f(Z)|, \tag{2.1}$$

where $\mathcal{F}$ is a class of real valued functions. Such a quantity is called Maximum Mean Discrepancy (MMD). Regardless of $\mathcal{F}$, (2.1) only defines a pseudo-metric on probability distributions (see [17]). In particular it is shown $\Delta(\cdot, \cdot)$ becomes a metric if $\mathcal{F} = H(k)$ is the reproducing kernel Hilbert space (RKHS) [17] associated with a kernel $k = k(\cdot, \cdot)$ that is *characteristic*.

Definition 2.2. *(Characteristic kernel)*
Let $\mathcal{F} = H(k)$ in (2.1) for some kernel k. Then k is a characteristic kernel if $\Delta(P, P_0) = 0$ implies $P = P_0$.

In practice the MMD has to be easily computed although the supremum in (2.1). One major interest of taking $\mathcal{F}$ as the unit ball of $H(k)$ is that $\Delta(P, P_0)$ can be cast as an easy to compute quantity as follows. Let us first introduce the Hilbert space embedding of a distribution P.

Definition 2.3. *(Hilbert space embedding, Lemma 3 from [7]) Let P be a distribution such that $\mathbb{E}_{Y \sim P} \sqrt{k(Y, Y)} < +\infty$.*
Then there exists $\mu_P \in H(k)$ such that for every $f \in H(k)$,

$$< \mu_P, f >= \mathbb{E} f(Y) \ . \tag{2.2}$$

μ_P is called the Hilbert space embedding of P in $H(k)$.

Then $\Delta(P, P_0)$ can be expressed as the gap between the Hilbert space embeddings of P and P_0:

$$\begin{aligned} \Delta(P, P_0) &= \sup_{f \in H(k), ||f|| \leq 1} |\mathbb{E}_P f(Y) - \mathbb{E}_{P_0} f(Z)| \\ &= \sup_{f \in H(k), ||f|| \leq 1} | < \mu_P - \mu_{P_0}, f > | \\ &= ||\mu_P - \mu_{P_0}|| \ . \end{aligned} \tag{2.3}$$

Since μ_P can be estimated by $1/n \sum_{i=1}^n k(Y_i, \cdot)$, [6], an estimator of (2.3) can be derived as

$$\widehat{\Delta} = \frac{1}{n} \left(\sum_{i,j=1}^{n} [\, k(Y_i, Y_j) + k(Z_i, Z_j) - 2k(Y_i, Z_j) \,] \right)^{1/2} ,$$

where $(Y_1, \ldots, Y_n)$ and $(Z_1, \ldots, Z_n)$ are samples of *i.i.d.* random variables with respective distributions P and P_0.

However the MMD-based approach suffers two main drawbacks: (i) it requires a characteristic kernel, which restricts its applicability, and (ii) the distribution of the test statistic $\widehat{\Delta}$ has to be approximated at two levels, which reduces the statistical test power. On the one hand, one purpose of the present work is to design a strategy allowing to deal with very general objects. It is typically the setting where no characteristic kernel does necessarily exist, or at least where conditions to check the characteristic property are completely awkward (see [17]). On the other hand, the distribution of $\widehat{\Delta}$ is first approximated by its asymptotic one, which is an infinite sum of weighted non-centered chi-squares [7, Theorem 12]. Second the distribution parameters have to be approximated through the eigendecomposition of a recentered Gram matrix (see Section 3.2 in [8]), which is computationally costly.

3 New normality test in RKHS

3.1 Goal

Let $X_1, \ldots, X_n \in \mathcal{X}$ be *i.i.d.* random variables. One only require $\mathcal{X}$ can be equipped with a positive definite kernel k associated with $H(k)$. The typical example of $\mathcal{X}$ one may consider is a set of DNA sequences.

Focusing on $Y_i = k(X_i, .) \in H(k)$ for all $1 \leq i \leq n$, our goal is to test whether $Y_i = k(X_i, .)$ follows a Gaussian distribution $P_0 = \mathcal{N}(\mu, \Sigma)$. Since $H(k)$ is a function space, a Gaussian variable $Z \in H(k)$ is a *Gaussian process.*

Definition 3.1. *(Gaussian process)*
Z is a Gaussian process if there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for any $a_1, \ldots, a_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$, $\sum_{i=1}^n a_i Y(x_i)$ is a univariate Gaussian random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.
Its mean $\mu \in H(k)$ and covariance function $\Sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are defined for every $x, y \in \mathcal{X}$ by:

$$\mu(x) = \mathbb{E}Z(x), \qquad \Sigma(x, y) = \mathrm{cov}(Z(x), Z(y)) \ .$$

By considering $H(k)$ as a linear space instead of a function space, a Gaussian process can be defined in an equivalent way.

Definition 3.2. *(Gaussian process)*
Z is a Gaussian process if there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for any

$f \in H(k)$, $< Z, f >$ is a univariate Gaussian random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.
Its mean $\mu \in H(k)$ and covariance operator $\Sigma_{op} \in HS(H(k))$ are defined for every $f, g \in H(k)$ by:

$$\begin{aligned} < \mu, f >= & \ \mathbb{E} < Z, f >, \\ < \Sigma_{op} f, g >= & \ \mathrm{cov}(< Z, f >, < Z, g >), \end{aligned}$$

where $HS(H(k))$ denotes the space of all linear applications $H(k) \to H(k)$ with finite trace (Hilbert-Schmidt operators).

The means in Definitions 3.1 and 3.2 coincide. Σ and Σ_{op} are linked by the following equality for every $x, y \in \mathcal{X}$

$$< \Sigma_{op} k(x, .), k(y, .) >= \Sigma(x, y) \ .$$

Remark that if $\mathcal{X} = \mathbb{R}^d$ and $k =< ., . >_{\mathbb{R}^d}$, then $H(k)$ is the dual of $\mathbb{R}^d$ (that is the set of all linear forms $< x, . >_{\mathbb{R}^d}$ on $\mathbb{R}^d$). In this case, $H(k)$ is isomorphic to $\mathbb{R}^d$ and Gaussian processes in $H(k)$ are reduced to multivariate Gaussian variables in $\mathbb{R}^d$.

3.2 New test procedure

Let us assume μ and Σ are known, and also $\mathbb{E}Y_i = \mu = 0$ for every $1 \leq i \leq n$, for the sake of simplicity.

3.2.1 Algorithm

We provide the main steps of the whole test procedure, which are further detailed in Sections 3.2.2–3.2.4.

1. **Input:** $X_1, \ldots, X_n \in \mathcal{X}$, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (kernel), Σ (covariance function), and $0 < \alpha < 1$ (test level).
2. Compute $K = [\, k(X_i, X_j) \,]_{i,j}$ (Gram matrix) and $C = [\, \Sigma(X_i, X_j) \,]_{i,j}$ (covariance matrix).
3. Compute $n\hat{L}^2$ (test statistic) from (3.6) that depends on K and C (Section 3.2.3)
4. (a) Draw B Monte-Carlo samples $X_1^b, \ldots, X_n^b$ under H_0, for $b = 1, \ldots, B$.
   (b) Compute $\hat{q}_{\alpha,n}$ ($1 - \alpha$ quantile of $n\hat{L}^2$ under H_0) (Section 3.2.4).
5. **Output:** Reject H_0 if $n\hat{L}^2 > \hat{q}_{\alpha,n}$, and accept otherwise.

The computation time of $\hat{q}_{\alpha,n}$ is of order $\mathcal{O}(Bn^2)$, which is faster than estimating the MMD limit distribution quantile as long as $n \geq B$ (Section 3.2.4 and Section 5.3).

3.2.2 Laplace-MMD (L-MMD)

The Laplace-MMD test (L-MMD) follows the same idea as the MMD test, but improves upon it by relaxing the restrictive assumption of *characteristic kernel*. Using that Laplace transform $\mathcal{L}_U(t) = \mathbb{E}_U \exp(tU)$ characterizes the distribution of a random variable $U \in \mathbb{R}$, the gap between two distributions P and P_0 can be evaluated by

$$\begin{aligned}\Delta L &= \sup_{f \in H(k), ||f|| \leq 1} \left| \mathbb{E}_{Y \sim P} e^{<Y,f>} - \mathbb{E}_{Z \sim P_0} e^{<Z,f>} \right| \\ &= \sup_{||f||=1} \sup_{|t| \leq 1} |\mathcal{L}_{<Y,f>}(t) - \mathcal{L}_{<Z,f>}(t)| \ , \end{aligned} \tag{3.4}$$

where $< \cdot, \cdot >$ denotes the inner-product in $H(k)$. Therefore $\Delta L = 0$ implies $< Y, f >$ and $< Z, f >$ have the same distribution for every f, which provides $P = P_0$ by the Cramer-Wold theorem (Proposition 2.2).

Deriving an efficient test procedure requires to provide a quantity related to (3.4) that is easy to compute. Following (2.3) this is done rephrasing $\exp(< \cdot, \cdot >) = \bar{k}$ as a new positive definite kernel associated with a new RKHS $H(\bar{k})$. Thus it allows to get a computable form of (3.4).

Theorem 3.3. *Assume* $\max\left(\mathbb{E}_P e^{||Y||}, \mathbb{E}_{P_0} e^{||Z||}\right) < \infty$. *Let* $\bar{\mu}_P, \bar{\mu}_{P_0} \in H(\bar{k})$ *be respective embeddings of* P *and* P_0. *Then,*

$$L = L(P, P_0) := ||\bar{\mu}_P - \bar{\mu}_{P_0}||_{H(\bar{k})} \ , \tag{3.5}$$

equals zero if and only if $P = P_0$.

Proof. Introducing $\bar{k} = \exp(< \cdot, \cdot >)$, it comes

$$\begin{aligned}\Delta L &\leq \sup_{h \in H(\bar{k}), ||h|| \leq e^{1/2}} | < \bar{\mu}_P - \bar{\mu}_{P_0}, h > |_{H(\bar{k})} \\ &= e^{1/2} ||\bar{\mu}_P - \bar{\mu}_{P_0}||_{H(\bar{k})} = \ e^{1/2} L(P, P_0) \ , \end{aligned}$$

where the inequality results from $\left\{\bar{k}(f, \cdot), ||f||_{H(k)} \leq 1\right\} \subset \left\{h \in H(\bar{k}), ||h||_{H(\bar{k})} \leq e^{1/2}\right\}$. Therefore $L(P, P_0) = 0$ implies $\Delta L = 0$ and $P = P_0$. Conversely $P = P_0$ implies $\bar{\mu}_P = \bar{\mu}_{P_0}$ and $L(P, P_0) = 0$. □

3.2.3 New test statistic

As in [6], L can be estimated by replacing $\bar{\mu}_P$ with the sample mean $\hat{\bar{\mu}}_P = 1/n \sum_{i=1}^n e^{<Y_i, \cdot>}$, leading to Proposition 3.4.

Proposition 3.4. *Assume the null-distribution* P_0 *is Gaussian* $\mathcal{N}(0, \Sigma)$ *and the largest eigenvalue* λ *of* Σ *is smaller than* 1. *Then the following statistic* $n\hat{L}^2$ *is an unbiaised estimator of* nL^2, *with*

$$n\hat{L}^2 = \frac{1}{n-1} \sum_{i \neq j}^n e^{k(X_i, X_j)} - 2 \sum_{i=1}^n e^{\frac{1}{2}\Sigma(X_i, X_i)} + nb^2 \ , \tag{3.6}$$

where $b^2 := ||\bar{\mu}_{P_0}||^2 = \left[\det(I - \Sigma^2)\right]^{-1/2}$.

The eigenvalue condition $\lambda < 1$ is not restrictive. With any $\gamma > 0$ such that $\gamma\lambda < 1$, one can compare $\gamma^{1/2} Y_i$ with $\mathcal{N}(0, \gamma\Sigma)$. The Gram matrix becomes $K' = \gamma K$ and the covariance matrix $C' = \gamma^2 C$.

Since it involves $n \times n$ matrices, the computation time for $n\hat{L}^2$ is the same as that of $\widehat{\Delta}$ (Section 2.3), that is of order $\mathcal{O}(n^2)$.

Proof. We prove that (3.6) is an unbiaised estimator of nL^2, that is its mean equals nL^2.

$$\begin{aligned}
\mathbb{E}n\hat{L}^2 =& \frac{1}{n-1}\sum_{i\neq j}^{n} \mathbb{E}e^{k(X_i,X_j)} - 2\sum_{i=1}^{n} \mathbb{E}e^{\frac{1}{2}\Sigma(X_i,X_i)} + nb^2 \\
=& \frac{1}{n-1}\sum_{i\neq j}^{n} \mathbb{E}e^{<Y_i,Y_j>_{H(k)}} \\
& - 2\sum_{i=1}^{n} \mathbb{E}_{Z\sim P_0} e^{<\Sigma_{op}Y_i,Y_i>_{H(k)}} + n||\mu_{P_0}||^2 \\
=& n||\mu_P||^2_{H(\bar{k})} - 2n < \mu_P, \mu_{P_0} >_{H(\bar{k})} + ||\mu_{P_0}||^2_{H(\bar{k})} \\
=& n||\mu_P - \mu_{P_0}||^2 = nL^2 \ .
\end{aligned}$$

□

3.2.4 Quantile estimation

Designing a test with confidence level $0 < \alpha < 1$ requires to compute the smallest $\epsilon > 0$ such that $\mathbb{P}_{H_0}(n\hat{L}^2 > \epsilon) \leq \alpha$ (Type-I error), which is the $1-\alpha$ quantile of the $n\hat{L}^2$ distribution under H_0 denoted by $q_{\alpha,n}$. Unfortunately $q_{\alpha,n}$ is unknown and has to be estimated.

Our purpose is to improve on the MMD strategy described in [7] in terms of power of detection by considering the *finite sample* null-distribution of $n\hat{L}^2$ rather than the asymptotic one. The improvement allowed by our strategy is illustrated by empirical results (see Figure 4 for instance).

Since the H_0-distribution $P_0 = \mathcal{N}(0, \Sigma)$ is known, $B > 0$ *i.i.d.* copies $n\hat{L}^2_{(1)}, \ldots, n\hat{L}^2_{(B)}$ of $n\hat{L}^2$ are drawn to estimate $q_{\alpha,n}$. More precisely for each $1 \leq b \leq B$,

$$\begin{aligned}
n\hat{L}^2_{(b)} = & \frac{1}{n-1}\sum_{i\neq j} e^{<Z_i^{(b)}, Z_j^{(b)}>} + n||\bar{\mu}_{P_0}||^2 \\
& - 2\sum_{i=1}^{n} \exp\left(\frac{1}{2} < Z_i^{(b)}, \Sigma_{op} Z_i^{(b)} >\right) \ , \qquad (3.7)
\end{aligned}$$

where $Z_1^{(b)}, \ldots, Z_n^{(b)} \overset{i.i.d.}{\sim} P_0$. Let us consider covariance Σ_{op} with a finite eigenvalue decomposition $\Sigma_{op} = \sum_{r=1}^{d} \lambda_r \Psi_r^{\otimes 2}$, with nonincreasing eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$

and eigenvectors $\left\{\Psi_i^{\otimes 2}\right\}_{i=1,\dots,d}$. From $G_{i,r}^{(k)} := \lambda_r^{-1/2} < Z_i^k, \Psi_r >$, the $(G_{i,r}^{(k)})_{i,r,k}$s are independent real-valued $\mathcal{N}(0,1)$, which leads to

$$< Z_i^{(k)}, Z_j^{(k)} > = \sum_{r=1}^{d} \lambda_r G_{i,r}^{(k)} G_{j,r}^{(k)}$$

$$< Z_i^{(k)}, \Sigma_{op} Z_i^{(k)} > = \sum_{r=1}^{d} \lambda_r^2 \left[G_{i,r}^{(k)}\right]^2 \ .$$

Let us now explain how the quantile estimator is computed. Assuming these B copies of $n\hat{L}^2$ are ordered in increasing order $n\hat{L}^2_{(1)} \leq \cdots \leq n\hat{L}^2_{(B)}$, let us define

$$\hat{q}_{\alpha,n} := n\hat{L}^2_{(\ell)}, \qquad \ell = \lfloor B + 2 - \alpha(B+1) \rfloor \tag{3.8}$$

where $\lfloor \cdot \rfloor$ denotes the integer part. This particular choice of ℓ is completely justified by the Type-I error control provided in Proposition 4.1. Finally the rejection region is defined by

$$\mathcal{R}_\alpha = \{ n\hat{L}^2 > \hat{q}_{\alpha,n} \} \ . \tag{3.9}$$

Estimating $q_{\alpha,n}$ requires simulating $B \times n \times d$ real Gaussian variables $\mathcal{N}(0,1)$ and computing B copies of $n\hat{L}^2$. Since with only n observations assuming $d > n$ seems unrealistic, the overall computational complexity is of order $\mathcal{O}(Bn^2)$. Note that the MMD quantile estimation proposed in [7] involves the computation of the eigenvalue decomposition of $n \times n$ matrices, which has a complexity bounded by $\mathcal{O}(n^3 + (n\log^2(n))\log(b))$, where the precision is of order 2^{-b} [14]. Then our strategy is preferable as long as n is large enough with respect to B, which is illustrated by Figure 5.

4 Theoretical assessment

4.1 Type-I error

The estimator of q_α defined by (3.8) depends on the ℓ-th ordered statistic $n\hat{L}^2_{(\ell)}$, where $\ell = \lfloor B + 2 - \alpha(B+1) \rfloor$. The purpose of the following result is to justify this somewhat unintuitive choice for ℓ by considering the Type-I error of the resulting procedure.

Proposition 4.1. *(Type-I error)*
Assume $P = P_0$ and $\alpha \geq 1/(B+1)$. With $\hat{q}_{\alpha,n}$ given by (3.8), it comes

$$\alpha - \frac{1}{B+1} \leq \mathbb{P}(n\hat{L}^2 > \hat{q}_{\alpha,n}) \leq \alpha \ . \tag{4.10}$$

Sketch of proof. The proof is straightforwardly derived from the cumulative function of the order statistic $\hat{q}_{\alpha,n}$, the density of a Beta distribution and the bounds $(1-\alpha)(B+1) \leq \ell \leq B + 2 - \alpha(B+1)$. □

Note that for a user-specified level $0 < \alpha < 1$, the L-MMD procedure requires to draw $B \geq 1/\alpha - 1$ samples to compute $\hat{q}_{\alpha,n}$. Besides the upper bound on the Type-I error is tight since the discrepancy between lower and upper bounds is not larger than $1/(B+1)$, which can be made negligible.

4.2 Type-II error

We now assume $P \neq P_0$. Theorem 4.2 gives the magnitude of the Type-II error, that is the probability of wrongly accepting H_0.

Before stating Theorem 4.2, let us introduce or recall useful notation.

- $L = ||\bar{\mu}_P - \bar{\mu}_{P_0}||$
- $q_{\alpha,n}$ is the $(1-\alpha)$-quantile of $n\hat{L}^2$ under the null-hypothesis
- Let $m_P^{(2)} = \mathbb{E}_P||\bar{\phi}(Y) - \bar{\mu}_P||^2$

Since $n\hat{L}^2$ converges weakly to a sum of weighted chi-squares (see [16], p. 194), $q_{\alpha,n}$ is close to a constant when $n \to +\infty$. L and $m_P^{(2)}$ do not depend on n.

The proof for Theorem 4.2 is provided in Appendix A.

Theorem 4.2. *(Type II error)*
Assume $||Y|| \leq M$ *(P-almost surely) for some* $0 < M < +\infty$.
Then, for any $n > (q_{\alpha,n} + m_P^{(2)})L^{-2}$

$$\mathbb{P}(n\hat{L}^2 \leq \hat{q}_{\alpha,n}) \leq \exp\left(-\frac{n\left\{L - \sqrt{(q_{\alpha,n} + m_P^{(2)})/(n-1)}\right\}^2}{f_1(n) + f_2(M,L,n)}\right) f_3(B,M,L) \ , \qquad (4.11)$$

where

$$f_1(n) = 2m_P^{(2)} + \mathcal{O}_n(1/\sqrt{n})$$

$$f_2(M,L,n) = \left\{\frac{8\sqrt{2}}{3}L^2\exp(M^2/2) + L\mathcal{O}_n(1/n)\right\}\left(1 + \mathcal{O}_n(1/\sqrt{n})\right) f_1^{1/2}(n)$$

$$f_3(B,M,L) = 1 + \frac{3C_{P_0}}{8\exp(M^2/2)L^2\sqrt{2m_P^{(2)}\alpha B}} + \frac{o_B(1/\sqrt{B})}{\exp(M^2)L^4} \ ,$$

where C_{P_0} *only depends on* P_0 *and the "*$\mathcal{O}_n$*" and "*o_B*" terms are idnependent of* L *and* M.

The upper bound in (4.11) shows an exponential decrease for the Type-II error when n grows. Furthermore, it reflects the expected behaviour of the Type-II error with respect to meaningful quantities

- When L decreases, the bound increases which is relevant as the alternative becomes more difficult to detect,
- When M gets smaller, the departure between P_0 and P is widened and as a result the upper bound decreases,
- When α (Type-I error) decreases, $q_{\alpha,n}$ gets larger and so does the bound.

Remark that the assumption $||Y|| \leq M$ P-a.s. is fulfilled if a bounded kernel k is considered.

5 Experiments

5.1 Type-I/II errors study

Empirical performances of L-MMD are inferred on the basis of synthetic data. L-MMD is compared with two other procedures: Random Projection (Section 2.2) and the asymptotic version of L-MMD denoted "L-MMDa".

We set $\mathcal{X} = \mathbb{R}^d$ and $k =< .,. >_{\mathbb{R}^d}$ (where $d = 25$) and thus $H(k)$ is reduced to $\mathbb{R}^d$. Therefore, L-MMD is used as a multivariate normality test and synthetic data are d-dimensional observations drawn from a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.

To control the difficulty level in the experiments, we introduced two parameters $\delta, \lambda \geq 0$ such that $\mu = \delta \cdot (1, 1/2, \ldots, 1/d)'$, and $\Sigma = \lambda \cdot \text{diag}(1, 1/4, \ldots, 1/d^2)$, where $\text{diag}(u)$ denotes the diagonal matrix with diagonal equal to $u \in \mathbb{R}^d$. For the Random Projection test, data are projected onto a randomly chosen direction generated from a zero-mean Gaussian distribution of covariance $\text{diag}(1, \ldots, d^{-2})$.

5.1.1 Type-I error

The left panel of Figure 1 displays the Type-I error of L-MMDa and L-MMD with respect to B. Indeed B independent samples from the asymptotic distribution of $n\hat{L}^2$ have been

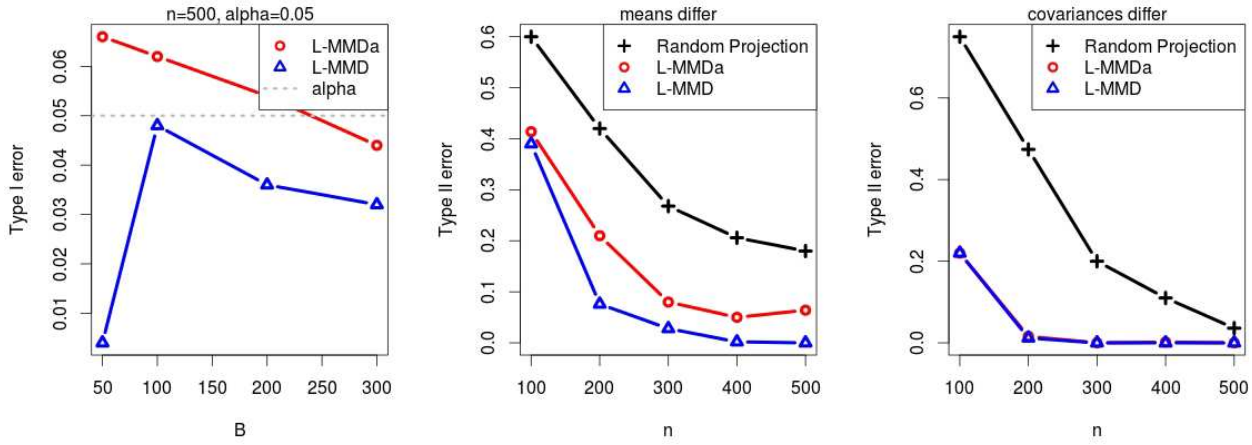

Figure 1: **Left:** Type-I errors of the L-MMDa (• red) and L-MMD (Δ blue) tests. **Center-Right:** Type-II errors of the Random Projection (+ black), L-MMDa (• red) and L-MMD (Δ blue) tests. Center: The null-distribution and alternative means differ. Right: The null-distribution and the alternative covariances differ. A theoretical prevision of the L-MMD Type-II error is also plotted (dashed purple).

drawn to allow the comparison between L-MMDa and L-MMD. Random Projection is not included since it does not depend on B samples. Observations are generated from the null-distribution with $\delta_0 = 0$ and $\lambda_0 = 0.5$. The test level is $\alpha = 0.05$ and $n = 500$. B ranges from 50 to 300. 500 simulations are performed for each B and each test.

The Type-I error of L-MMDa is always larger than that of L-MMD although the gap between them remains small (≤ 0.01) for $B \geq 100$. L-MMD always remains below the prescribed test level α unlike L-MMDa for $B \leq 250$.

5.1.2 Type-II error

The same P_0 (null-distribution) as in Section 5.1.1 is used and two alternatives are considered. The first one differs from P_0 by the mean ($\delta_{A1} = 0.15$ for the alternative). The second one has the same mean as P_0 but a different covariance ($\lambda_{A2} = 0.75\lambda_0$). Results are displayed in the center (different means) and right (different covariances) of Figure 1. We also plotted the prevision of the L-MMD performance provided by Theorem 4.2.

L-MMDa and L-MMD both outperform Random Projection that shows the worst overall performance. As n grows, L-MMD seems more powerful than L-MMDa ($n \geq 200$).

5.2 Influence of the dimensionality

One main concern of goodness-of-fit tests is their drastic loss of power as dimensionality increases. Empirical evidences (see Table 3 in [19]) prove ongoing multivariate normality tests suffer such deficiencies. The purpose of the present section is to check if the good behavior of L-MMD (observed in Section 5.1 when $d = 25$) stills holds in high or infinite dimension.

In Section 5.2.1, two different settings ($d = 2$ and $d = 25$) are explored with synthetic data where the L-MMD performance is compared with that of two goodness-of-fit tests (Henze-Zirkler and Energy Distance). Real data serve as infinite dimensional setting in Section 5.2.2 to assess the L-MMD power.

5.2.1 Finite-dimensional case (Synthetic data)

The power of our test is compared with that of two multivariate normality tests: the HZ test [9] and the energy distance test [19]. In what follows, we briefly recall the main idea of these tests.

The HZ test relies on the following statistic

$$HZ = \int_{\mathbb{R}^d} \left| \hat{\Psi}(t) - \Psi(t) \right|^2 \omega(t) dt \ , \tag{5.12}$$

where $\Psi(t)$ denotes the characteristic function of P_0, $\hat{\Psi}(t) = n^{-1} \sum_{j=1}^{n} e^{i<t,Y_j>}$ is the empirical characteristic function of the sample $Y_1, \ldots, Y_n$, and $\omega(t) = (2\pi\beta)^{-d/2} \exp(-||t||^2/(2\beta))$ with $\beta = 2^{-1/2}[(2d+1)n)/4]^{1/(d+4)}$. The H_0-hypothesis is rejected for large values of HZ.

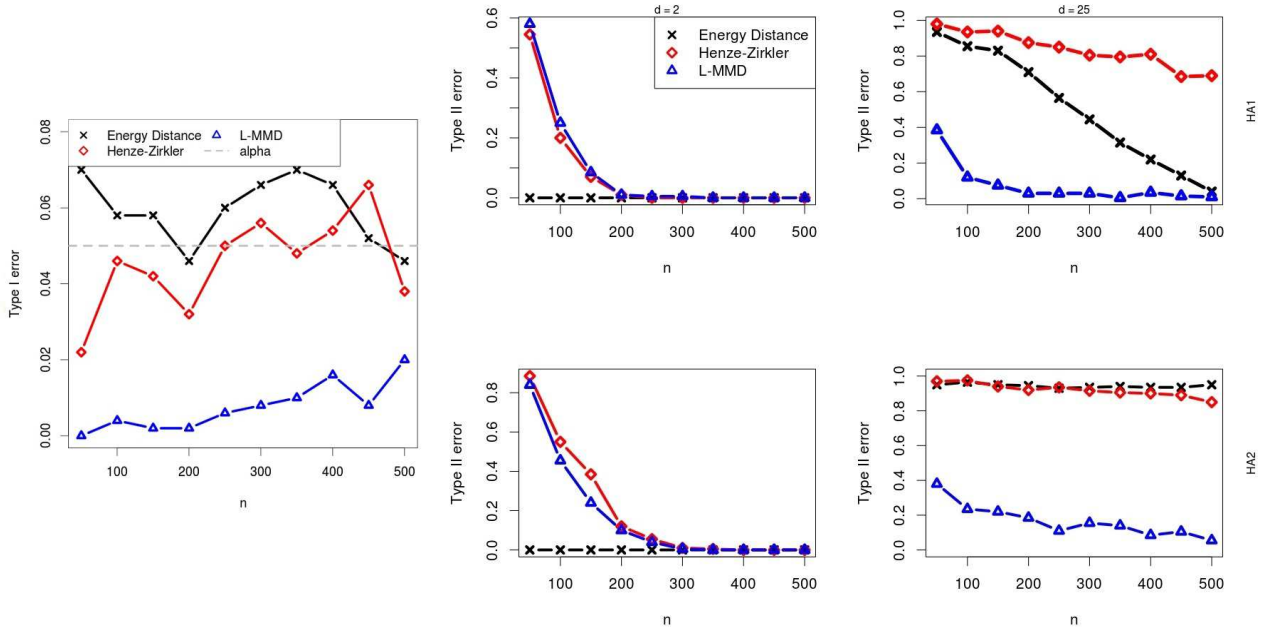

Figure 2: Type-I and type-II errors of L-MMD (Δ blue), Energy Distance ($\times$ black), and Henze-Zirkler ($\bullet$ red). For the Type-II error, two alternative distributions are considered: HA1 (top panel) and HA2 (bottom panel). Two settings are considered: $d = 2$ (left) and $d = 25$ (right).

The energy distance (ED) test is based on

$$\mathcal{E}(P, P_0) = 2\mathbb{E}||Y - Z||^2 - \mathbb{E}||Y - Y'||^2 - \mathbb{E}||Z - Z'||^2 \tag{5.13}$$

which is called the *energy distance*, where $Y, Y' \sim P$ and $Z, Z' \sim P_0$. Note that $\mathcal{E}(P, P_0) = 0$ if and only if $P = P_0$. The test statistic is given by

$$\begin{aligned}\hat{\mathcal{E}} =& \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_Z||Y_i - Z||^2 - \mathbb{E}_{Z,Z'}||Z - Z'||^2 \\ & - \frac{1}{n^2}\sum_{i,j=1}^{n} ||Y_i - Y_j||^2 \ ,\end{aligned} \tag{5.14}$$

where $Z, Z' \overset{i.i.d.}{\sim} P_0$ (null-distribution). HZ and ED tests set the H_0-distribution at $P_0 = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$ where $\hat{\mu}$ and $\hat{\Sigma}$ are respectively the standard empirical mean and covariance. Therefore, we consider the same null-hypothesis for the L-MMD.

Two alternatives are considered. A mixture of two Gaussians with different means ($\mu_1 = 0$ and $\mu_2 = 1.5\ (1, 1/2, \ldots, 1/d)$) and same covariance $\Sigma =$

$0.5\ \mathrm{diag}(1, 1/4, \ldots, 1/d^2)$, whose mixture proportions equals either $(0.5, 0.5)$ (alternative HA1) or $(0.8, 0.2)$ (alternative HA2).

200 simulations are performed for each test, each alternative and each n (ranging from 100 to 500). B is set at $B = 250$ for L-MMD.

The test level is set at $\alpha = 0.05$ for all tests. Since empirical parameters are considered in all tests, the actual Type-I error may not be controlled anymore. The left plot in Figure 2 confirms that the actual Type-I error for HZ and ED tests remain more or less around α (± 0.03). The Type-I error for L-MMD is still upper bounded by α and gets closer to the prescribed test level as n increases.

As for the Type-II error, experimental results (Figure 2) reveal two different behaviors as d increases (from center to right columns). Whereas both HZ and ED tests lose power, L-MMD still exhibits similar Type-II error values. The same conclusion holds true under HA1 and HA2 as well, while the failure of HZ and ED is stronger with HA2 (more difficult). This confirms that HZ and ED tests are not suited to high-dimensional settings unlike L-MMD. Notice that when d is small, L-MMD and HZ have almost the same Type-II error. This can be due to the integration involved in the HZ statistic. As d increases any

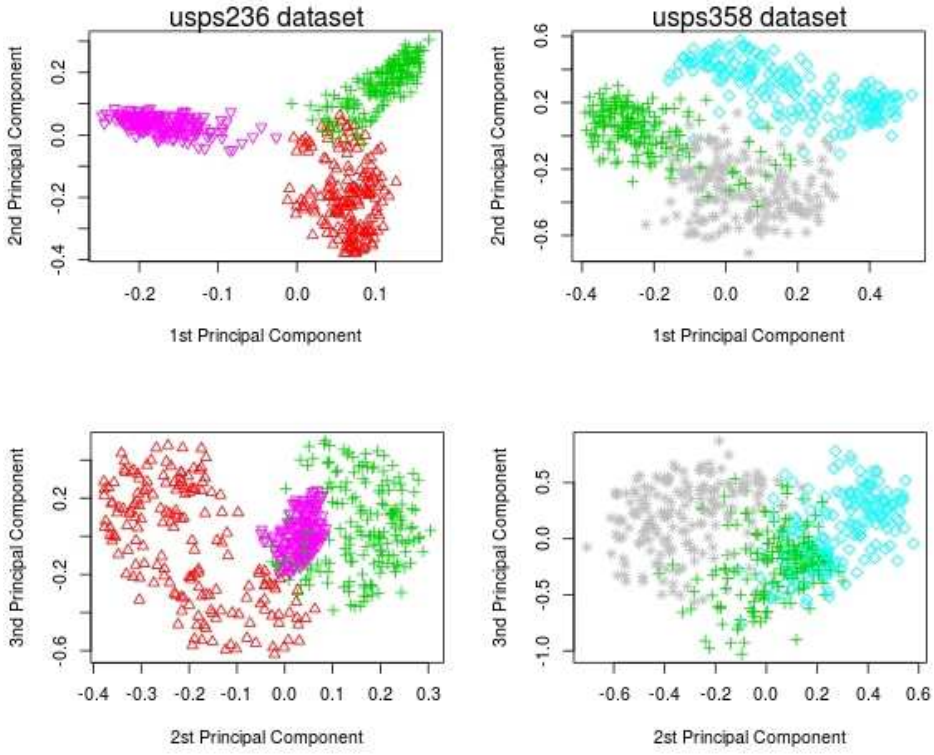

Figure 3: 3D-Visualization (Kernel PCA) of the "Usps236" (left) and "Usps358" (right) datasets

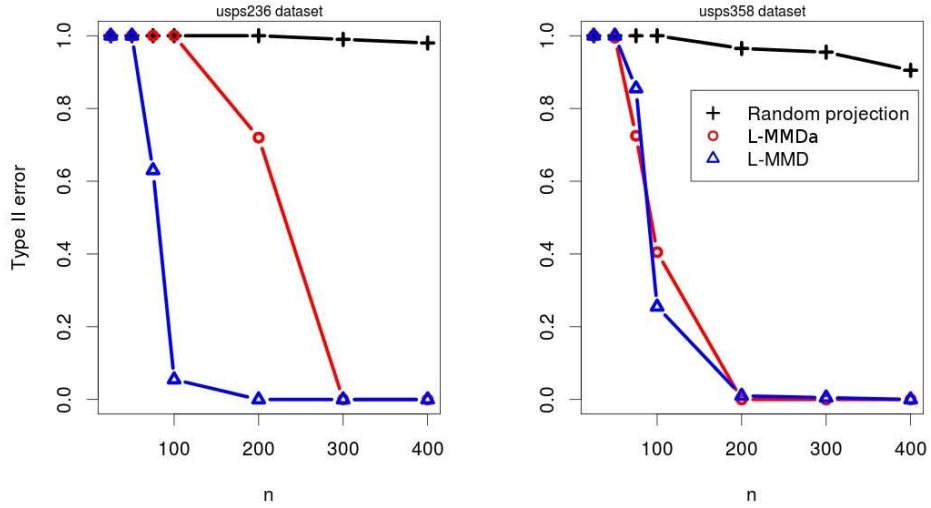

Figure 4: Comparison of Type-II error for: L-MMD (Δ blue), L-MMDa ($\bullet$ red) and Random Projection (+ black). Left: "Usps236". Right: "Usps358".

discrepancy arising in only a few dimensions is neglected in front of the leading behavior in all other directions. On the contrary, the supremum at the core of L-MMD (3.4) takes into account this kind of discrepancy.

5.2.2 Infinite-dimensional case (real data)

Let us consider the USPS dataset (UCI machine learning repository: *http://archive.ics.uci.edu*), which consists of handwritten digits, split up into 10 classes (each for a digit). Each observation represents a 8×8 greyscale matrix as a 64-dimensional vector. A Gaussian kernel $k_G(\cdot,\cdot) = \exp(-\sigma^2||\cdot - \cdot||^2)$ is used with $\sigma^2 = 10^{-4}$. Data are visualized through a Kernel PCA [15] and displayed in Figure 3. Comparing sub-datasets "Usps236" (keeping the three classes "2", "3" adn "6", 541 observations) and "Usps358" (classes "3", "5" and "8", 539 observations), the 3D-visualization suggests three well-separated Gaussian components for "Usps236" (left panels), and more overlapping classes for "Usps358" (right panels). Therefore from these two non-Gaussian settings, the last one seems more difficult to detect.

As in Section 5.1 our test is compared with Random Projection (RP) and L-MMDa tests, specially designed for infinite-dimensional settings. For RP, a univariate Kolmogorov-Smirnov test is performed from the projection onto a randomly chosen direction generated by a zero-mean Gaussian process of covariance k_G. The test level $\alpha = 0.05$ and 100 repetitions have been done for each sample size.

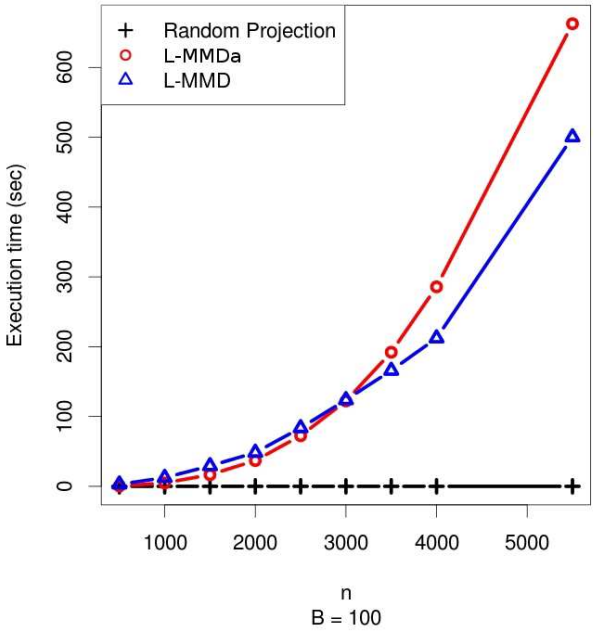

Figure 5: Execution time of L-MMD (Δ blue), L-MMDa ($\bullet$ red) and Random Projection ($+$ black).

Results in Figure 4 match those obtained in the finite-dimensional case (Figure 2). On the one hand, RP is by far less powerful than L-MMDa and L-MMD in both cases. Its Type-II error remains close to 1 whereas L-MMD always truly rejects H_0 for $n \geq 200$. On the other hand, L-MMD seems more powerful than L-MMDa with "Usps236" since it is close to 0 for $n \geq 100$ while L-MMDa reaches similar values only for $n \geq 300$. However both L-MMDa and L-MMD exhibit a similar behavior in terms of Type-II error with "Usps358", and always reject H_0 for $n \geq 200$. This may be due to the higher difficulty of this dataset that do not allow to clearly distinguish between test procedures.

5.3 Execution Time

From Sections 5.1 and 5.2 L-MMD is shown to outperform L-MMDa in terms of power. This may result from the asymptotic approximation underlying the L-MMDa procedure, while the L-MMD test is performed with the non-asymptotic distribution. The present section aims at verifying this gain in performance is not balanced by a larger computation time.

From the remark at the end of Section 3.2.4, L-MMD seems less computationally demanding than L-MMDa as long as n is large enough with respect to B. We carried

out an experiment with synthetic data where $B = 100$ and n ranges from 500 to 5500. No parallelization has been made in this experiment. From Figure 5 results support the above conclusion. For $n \leq 3000$, L-MMDa and L-MMD have similar computation time, L-MMDa being only slightly faster. However $n > 3000$ illustrates the predicted phenomenon. L-MMD is significantly less time consuming than L-MMDa. Since the L-MMDa execution time is of order $\mathcal{O}(n^3)$, the L-MMD complexity of order $\mathcal{O}(Bn^2)$ becomes smaller as the sample size increases.

6 Conclusion

We introduced a new normality test in RKHS. It turns out to be more powerful than ongoing high- or infinite-dimensional tests (such as random projection). In particular, empirical studies showed a mild sensibility to high-dimensionality for the L-MMD. Therefore L-MMD can be used as a multivariate normality (MVN) test without suffering a loss of power when d gets larger unlike other MVN tests (Henze-Zirkler, Energy-distance).

An aspect that most goodness-of-fit tests neglect is the estimation of the distribution parameters (here the mean and covariance of a Gaussian distribution). Indeed little is known about how much it affects the test performances. Adapting our test to this framework would be welcome in future investigations.

A Proof of Theorem 4.2

A.1 Main proof

The goal is to get an upper bound for the Type-II error

$$\mathbb{P}(n\hat{L}^2 \leq \hat{q} \mid \mathcal{H}_A) \ . \tag{A.15}$$

In the following, the feature map from $H(k)$ to $H(\bar{k})$ will be denoted as

$$\bar{\phi} : H(k) \to H(\bar{k}), \ \ y \mapsto \bar{k}(y, .) \ .$$

1. **Reduce $n\hat{L}^2$ to a sum of independent terms**
   The first step consists in getting a tight upper bound for (A.15) which involve a sum of independent terms. This will allow the use of a Bennett concentration inequality in the next step.
   $n\hat{L}^2$ is expanded as follows

$$\begin{aligned} n\hat{L}^2 =& \frac{1}{n-1} \sum_{i \neq j}^{n} < \bar{\phi}(Y_i) - \bar{\mu}_{P_0}, \bar{\phi}(Y_j) - \bar{\mu}_{P_0} > \\ :=& n\hat{L}_P^2 + nL^2 + 2nS_n \ . \end{aligned} \tag{A.16}$$

   where $\hat{L}_P^2 = [n(n-1)]^{-1} \sum_{i \neq j}^{n} < \bar{\phi}(Y_i) - \bar{\mu}_P, \bar{\phi}(Y_j) - \bar{\mu}_P >$ and $S_n = < \hat{\bar{\mu}}_P - \bar{\mu}_P, \bar{\mu}_P - \bar{\mu}_{P_0} >$ with $\hat{\mu}_P = n^{-1} \sum_{i=1}^{n} \bar{\phi}(Y_i)$.

It corresponds to the so-called Hoeffding expansion of the U-statistic $\hat{L}^2$ [10] written as a sum of degenerate U-statistics. Since $n\hat{L}_P^2$ converges weakly to a sum of weighted chi-squares and $\sqrt{n}S_n$ to a Gaussian, $\hat{L}_P^2$ becomes negligible with respect to S_n when n is large. Therefore, we consider a surrogate for the Type-II error (A.15) by removing $\hat{L}_P^2$ with a negligible loss of accuracy.

Using Lemma A.3, $\hat{L}_P^2$ can be split up into a non-negative quantity and a sum of independent variables

$$\hat{L}_P^2 = \frac{n}{n-1}||\hat{\bar{\mu}}_P - \bar{\mu}_P||^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}||\bar{\phi}(Y_i) - \bar{\mu}_P||^2 \ . \tag{A.17}$$

Writing (A.15) conditionally to $\hat{q}$, plugging (A.16) and (A.17) into (A.15) and using $||\hat{\bar{\mu}}_P - \bar{\mu}_P||^2 \geq 0$ yield the upper bound

$$\mathbb{P}(n\hat{L}^2 \leq \hat{q} \mid \hat{q}) \leq \mathbb{P}\left(-\frac{1}{n-1}\sum_{i=1}^{n}||\bar{\phi}(Y_i) - \bar{\mu}_P||^2 + nL^2 + 2nS_n \leq \hat{q}\|\hat{q}\right) \ . \tag{A.18}$$

Remark that both positive and negative terms of (A.17) are of the same order than $\hat{L}_P^2$ (that is of order n^{-1}) so that the loss of accuracy in the bound (A.18) is negligeable.

$$\mathbb{P}(n\hat{L}^2 \leq \hat{q} \mid \hat{q}) \ \leq \ \mathbb{P}(\sum_{i=1}^{n} f(Y_i) \geq n\hat{s} \mid \hat{q}) \ , \tag{A.19}$$

where

$$f(Y_i) := \ \frac{||\bar{\phi}(Y_i) - \bar{\mu}_P||^2}{n-1} - 2 < \bar{\phi}(Y_i) - \bar{\mu}_P, \bar{\mu}_P - \bar{\mu}_{P_0} > \ , \quad \hat{s} := \ L^2 - \frac{\hat{q}}{n} - \frac{m_P^{(2)}}{n-1} \ ,$$

and $m_P^{(i)} = \mathbb{E}||\phi(\bar{Y}_i) - \bar{\mu}_P||^i$ for any $i \geq 2$.

2. **Apply a concentration inequality**

We now want to find an upper bound for (A.19) through a concentration inequality, namely Lemma A.1 with $\xi_i = f(Y_i)$, $\epsilon = n\hat{s}$, $\nu^2 = \mathrm{Var}(f(Y_i))$ and $f(Y_i) \leq c = \bar{M}$ (P-almost surely).

Lemma A.1 combined with Lemma A.5 and A.4 yields the upper bound

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^{n} f(Y_i) \geq n\hat{s} \mid \hat{q}) &\leq \exp\left(-\frac{n\hat{s}^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta\hat{s}}\right)\mathbb{1}_{\hat{s}\geq 0} + \mathbb{1}_{\hat{s}<0} \\ &:= \exp(g(\hat{s}))\mathbb{1}_{\hat{s}\geq 0} + \mathbb{1}_{\hat{s}<0} := h(\hat{s}) \ ,\end{aligned} \tag{A.20}$$

where

$$\overline{M} := \left(4\sqrt{2}e^{M^2/2}L + \frac{m_P^{(2)}}{n-1}\right) \ , \qquad \vartheta^2 := \ L^2 m_P^{(2)} + \frac{L m_P^{(3)}}{n-1} + \frac{m_P^{(4)} - (m_P^{(2)})^2}{4(n-1)^2} \ .$$

3. **"Replace" the estimator $\hat{q}_{\alpha,n}$ with the true quantile $q_{\alpha,n}$ in the bound**

It remains to take the expectation with respect to $\hat{q}_{\alpha,n}$. In order to make it easy, $\hat{q}_{\alpha,n}$ is pull out of the exponential term of the bound. This is done through a Taylor-Lagrange expansion (Lemma A.6).

Lemma A.6 rewrites the bound in (A.20) as

$$\exp\left(-\frac{ns^2}{2\vartheta^2+(2/3)\overline{M}\vartheta s}\right)\left\{1+\frac{3n}{2\overline{M}\vartheta}\exp\left(\frac{3|\tilde{q}-q|}{2\overline{M}\vartheta}\right)\mathbb{1}_{\tilde{s}\geq 0}|\hat{s}-s|\right\} \ , \qquad \text{(A.21)}$$

where

$$s=L^2-\frac{q}{n}-\frac{b_P^{(2)}}{n-1} \ , \quad \tilde{s}=L^2-\frac{\tilde{q}}{n}-\frac{b_P^{(2)}}{n-1} \ , \quad \tilde{q}\in(q\wedge\hat{q},q\vee\hat{q}) \ ,$$

and $s\geq 0$ because of the assumption $n>(q+m_P^{(2)})L^{-2}$.

The mean (with respect to $\hat{q}$) of the right-side multiplicative term of (A.21) is bounded by

$$1+\frac{3n}{2\overline{M}\vartheta}\left\{\mathbb{E}_{\hat{q}}\left(\exp\left(\frac{3|\tilde{q}-q|}{\overline{M}\vartheta}\right)\mathbb{1}_{\tilde{s}\geq 0}\right)\right\}^{1/2}\sqrt{\mathbb{E}_{\hat{q}}(\hat{s}-s)^2} \ ,$$

because of the Cauchy-Schwarz inequality.

On one hand, using $\hat{q}\to q$ P_0-a.s. when $B\to+\infty$

$$\begin{aligned}\mathbb{E}_{\hat{q}}\left(\exp\left(\frac{3|\tilde{q}-q|}{\overline{M}\vartheta}\right)\mathbb{1}_{\tilde{s}\geq 0}\right) &= \mathbb{E}_{\hat{q}}\left(\left[1+\frac{o_B(|\hat{q}-q|)}{\overline{M}\vartheta}\right]\mathbb{1}_{\tilde{s}\geq 0}\right)\\ &\leq 1+\frac{\mathbb{E}_{\hat{q}}(o_B(|\hat{q}-q|)\mathbb{1}_{\tilde{s}\geq 0})}{\overline{M}\vartheta}= 1+\frac{o_B(1)}{\overline{M}\vartheta} \ , \qquad \text{(A.22)}\end{aligned}$$

which follows from the Dominated Convergence Theorem (since the variable $|\tilde{q}-q|\mathbb{1}_{\tilde{s}\geq 0}$ is bounded by the constant $|nL^2-q|\vee|q|$ for every B).

On the other hand, Lemma A.2 provides

$$\mathbb{E}(\hat{s}-s)^2=\frac{\mathbb{E}(\hat{q}-q)^2}{n^2}\leq\frac{C_{1,P_0}+\alpha C_{2,P_0}/B}{n^2\alpha B}\leq\frac{C_{P_0}}{n^2\alpha B} \ . \qquad \text{(A.23)}$$

so that an upper bound for the Type-II error is given by

$$\exp\left(-\frac{ns^2}{2\vartheta^2+(2/3)\overline{M}\vartheta s}\right)\left\{1+\frac{3C_{P_0}}{2\overline{M}\vartheta\sqrt{\alpha B}}+\frac{o_B(B^{-1/2})}{\overline{M}^2\vartheta^2}\right\} \ . \qquad \text{(A.24)}$$

Finally (A.24) can be bounded via the inequalities $n>(q+m_P^{(2)})/L^2$ and $\overline{M}\vartheta\geq 4\sqrt{2m_P^{(2)}}\exp(M^2/2)L^2$

$$\exp\left(-\frac{n\left[L-\sqrt{(q+m_P^{(2)})/(n-1)}\right]^2}{f_1(n)+f_2(M,L,n)}\right)f_3(B,M,L) \ ,$$

where

$$f_1(n) = 2m_P^{(2)} + \mathcal{O}_n(1/\sqrt{n})$$
$$f_2(M,L,n) = \left\{ \frac{8\sqrt{2}}{3} L^2 \exp(M^2/2) + L\mathcal{O}_n(1/n) \right\} \left(1 + \mathcal{O}_n(1/\sqrt{n})\right) f_1^{1/2}(n)$$
$$f_3(B,M,L) = 1 + \frac{3C_{P_0}}{8\exp(M^2/2)L^2\sqrt{2m_P^{(2)}\alpha B}} + \frac{o_B(1/\sqrt{B})}{\exp(M^2)L^4} \ .$$

Theorem 4.2 is proved.

A.2 Auxilary results

Lemma A.1. *(Bennett's inequality, Theorem 2.9 in [2]) Let $\xi_1, \ldots, \xi_n$ i.i.d. zero-mean variables bounded by c and of variance ν^2.*
Then, for any $\epsilon > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} \xi_i \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2n\nu^2 + 2c\nu\epsilon/3}\right) \ . \tag{A.25}$$

Lemma A.2. *Assume $\alpha < 1/2$. Then,*

$$\mathbb{E}(\hat{q}_{\alpha,n} - q_{\alpha,n})^2 \leq \frac{C_{1,P_0}}{\alpha B} + \frac{C_{2,P_0}}{B^2} \ , \tag{A.26}$$

where C_{1,P_0} and C_{2,P_0} only depends on P_0.

Proof. (Lemma A.2) Let $U_n = n\hat{L}^2_{\mathcal{H}_0,(\ell)}$ (under the null-hypothesis), $U_{n,1}, \ldots, U_{n,B}$ B i.i.d. copies of U_n, $U_{n,(1)} < \ldots < U_{n,(B)}$ the associated order statistics and q the $(1-\alpha)$-quantile of U_n, that is $\mathbb{P}(U_n > q) = \alpha$.
Condider $\ell = \lfloor B + 2 - \alpha(B+1) \rfloor$ and $\hat{q} := U_{n,(\ell)}$. $\mathbb{E}(\hat{q} - q)^2$ can be split up the following way

$$\mathbb{E}(\hat{q} - q)^2 = \mathrm{Var}(\hat{q}) + (\mathbb{E}\hat{q} - q)^2 \ . \tag{A.27}$$

Theorem 2.9. in [3] provides an upper bound for the variance term when $\ell \geq B/2$ (which holds since $\alpha < 1/2$)

$$\mathrm{Var}(\hat{q}) \leq \frac{2}{\alpha B} \mathbb{E}h^{-1}(U_{n,(\ell)}) \ ,$$

where h is the hazard rate of U_n defined by $h = f_n/(1 - F_n)$, F_n is the cumulative distribution function of U_n and $f_n = F_n^{'}$.

Since U_n converges weakly to a (possibly infinite) sum of weighted chi-squares (see [16], p. 194), $\mathbb{E}h^{-1}(U_{n,(\ell)})$ converges to a finite quantity as $n \to +\infty$. Therefore, there exists a quantity C_{P_0} which does not depend on n such that

$$\mathrm{Var}(\hat{q}) \leq \frac{C_{P_0}}{\alpha B} \ . \tag{A.28}$$

To bound the second additive term in (A.27), we determine which quantile of U_n $\mathbb{E}\hat{q}$ corresponds to.

$$\begin{aligned}\mathbb{P}(U_n \leq \mathbb{E}\hat{q}) =& \ \mathbb{E}_{\hat{q}}\mathbb{P}(U_n \leq \hat{q}|\hat{q}) = \mathbb{E}_{\hat{q}}\mathbb{E}_{U_n}\mathbb{1}_{U_n\leq\hat{q}} \\ =& \ \mathbb{E}_{U_n}\mathbb{E}_{\hat{q}}\mathbb{1}_{U_n\leq\hat{q}} = \mathbb{E}_{U_n}\mathbb{P}(\hat{q} \geq U_n|U_n) \ .\end{aligned}$$

The expression for the cdf of an order statistic yields

$$\mathbb{P}(U_n \leq \mathbb{E}\hat{q}) = \mathbb{E}_{U_n}\left\{\sum_{k=B-l}^{B}\binom{B}{k}(1-F_n(U_n))^k F_n^{B-k}(U_n)\right\} \ .$$

Since $F_n(U_n)$ follows a uniform distribution on $[0,1]$, the density of a Beta law appears in the latter equation, namely

$$\begin{aligned}\mathbb{P}(U_n \leq \mathbb{E}\hat{q}) =& \ \sum_{k=B-l}^{B}\binom{B}{k}\int_0^1 (1-u)^k u^{B-k}\mathrm{d}u \\ =& \ \sum_{k=B-l}^{B}\binom{B}{k}\frac{k!(B-k)!}{(B+1)!} = \frac{\ell+1}{B+1} \ .\end{aligned}$$

Let $Q_n = F_n^{-1}$ denote the quantile function of U_n. Since U_n converges to a sum of weighted chi-squares with quantile function Q_∞, one can write $Q_n = Q_\infty(1+o_n(1))$ (where $o_n(1)$ holds uniformly on the interval $[1-\alpha, (1-\alpha)+2/(\lceil 2/\alpha\rceil+1)]$). Besides, let $f_\infty = [Q_\infty^{-1}]'$ the density of the limit distribution of U_n.

By the Taylor-Lagrange expansion of Q_∞ of order 1, hence there exists $\xi \in (1-\alpha, (\ell+1)/(B+1))$ such that

$$\begin{aligned}|\mathbb{E}\hat{q} - q| =& \ \left|Q_\infty(\frac{\ell+1}{B+1}) - Q_\infty(1-\alpha)\right|(1+o_n(1)) \\ =& \ \left|Q_\infty(1-\alpha) + Q_\infty'(\xi)\left(\frac{\ell+1}{B+1} - (1-\alpha)\right) - Q_\infty(1-\alpha)\right|(1+o_n(1)) \\ \leq& \ \frac{2(1+o_n(1))}{(B+1)f_\infty(Q_\infty(\xi))} = \frac{2(1+o_n(1))(1+o_B(1))}{(B+1)f_\infty(q)} \leq \frac{C_{2,P_0}}{B} \ , \end{aligned} \tag{A.29}$$

where C_{2,P_0} only depends on P_0.
(A.27) combined with (A.28) and (A.29) yield the wanted bound.

Lemma A.3.

$$\hat{L}_P^2 = \frac{n}{n-1}||\hat{\bar{\mu}}_P - \bar{\mu}_P||^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}||\bar{\phi}(Y_i) - \bar{\mu}_P||^2 \ . \tag{A.30}$$

Proof.

$$\begin{aligned}
\hat{L}_P^2 =& \frac{1}{n(n-1)} \sum_{i \neq j}^{n} < \bar{\phi}(Y_i) - \bar{\mu}_P, \bar{\phi}(Y_j) - \bar{\mu}_P > \\
=& \frac{1}{n(n-1)} \sum_{i,j=1}^{n} < \bar{\phi}(Y_i) - \bar{\mu}_P, \bar{\phi}(Y_j) - \bar{\mu}_P > - \frac{1}{n(n-1)} \sum_{i=1}^{n} ||\bar{\phi}(Y_i) - \bar{\mu}_P||^2 \\
=& \frac{n}{n-1} < \frac{1}{n} \sum_{i=1}^{n} \bar{\phi}(Y_i) - \bar{\mu}_P, \frac{1}{n} \sum_{j=1}^{n} \bar{\phi}(Y_j) - \bar{\mu}_P > - \frac{1}{n(n-1)} \sum_{i=1}^{n} ||\bar{\phi}(Y_i) - \bar{\mu}_P||^2 \\
=& \frac{n}{n-1} ||\hat{\bar{\mu}}_P - \bar{\mu}_P||^2 - \frac{1}{n(n-1)} \sum_{i=1}^{n} ||\bar{\phi}(Y_i) - \bar{\mu}_P||^2 \ .
\end{aligned}$$

□

Lemma A.4. *If $||Y|| \leq M$ P-a.s., then $f(Y)$ is also bounded*

$$|f(Y)| \leq \overline{M} := 4\sqrt{2} e^{M^2/2} L + \frac{m_P^{(2)}}{n-1} \ . \tag{A.31}$$

Proof.

$$\begin{aligned}
|f(Y)| =& \left| 2 < \bar{\phi}(Y) - \bar{\mu}_P, \bar{\mu}_P - \bar{\mu}_{P_0} > - \frac{m_P^{(2)}}{n-1} \right| \\
\leq& \ 2||\bar{\phi}(Y) - \bar{\mu}_P|| L + \frac{m_P^{(2)}}{n-1} \\
\leq& \ 2\sqrt{2}||\bar{\phi}(Y)|| L + 2\sqrt{2}||\bar{\mu}_P|| L + \frac{m_P^{(2)}}{n-1} \\
=& \ 2\sqrt{2}\mathbb{E} e^{||Y||^2/2} L + 2\sqrt{2}\mathbb{E} e^{<Y,Y'>/2} L + \frac{m_P^{(2)}}{n-1} \\
\leq& \ 4\sqrt{2} e^{M^2/2} L + \frac{m_P^{(2)}}{n-1} := \overline{M} \ .
\end{aligned}$$

□

Lemma A.5.

$$\nu^2 \leq \vartheta^2 := L^2 m_P^{(2)} + \frac{L m_P^{(3)}}{n-1} + \frac{m_P^{(4)} - (m_P^{(2)})^2}{4(n-1)^2} \ . \tag{A.32}$$

Proof.

$$\begin{aligned}\nu^2 := \mathrm{Var}(g(Y)) = &\ \mathbb{E} < \bar{\phi}(Y_i) - \bar{\mu}_P, \bar{\mu}_P - \bar{\mu}_{P_0} >^2 + \frac{\mathbb{E}||\bar{\phi}(Y_i) - \bar{\mu}_P||^4}{4(n-1)^2} \\ &- \frac{\mathbb{E}(< \bar{\phi}(Y_i) - \bar{\mu}_P, \bar{\mu}_P - \bar{\mu}_{P_0} > ||\bar{\phi}(Y_i) - \bar{\mu}_P||^2)}{n-1} - \left[\frac{\mathbb{E}||\bar{\phi}(Y_i) - \bar{\mu}_P||^2}{2(n-1)}\right]^2 \\ \leq &\ L^2 m_P^{(2)} + \frac{L m_P^{(3)}}{n-1} + \frac{m_P^{(4)} - (m_P^{(2)})^2}{4(n-1)^2} := \vartheta^2 \ ,\end{aligned}$$

□

Lemma A.6. *Let h be defined as in* (A.20)*. Then,*

$$h(\hat{s}) \leq \exp\left(-\frac{ns^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta s}\right)\left\{1 + \frac{3n}{2\overline{M}\vartheta}\exp\left(\frac{3|\tilde{q} - q|}{2\overline{M}\vartheta}\right)\mathbb{1}_{\tilde{s}\geq 0}|\hat{s} - s|\right\} \ , \qquad \text{(A.33)}$$

where

$$s = L^2 - \frac{q}{n} - \frac{b_P^{(2)}}{n-1} \ , \quad \tilde{s} = L^2 - \frac{\tilde{q}}{n} - \frac{b_P^{(2)}}{n-1} \ , \quad \tilde{q} \in (q \wedge \hat{q}, q \vee \hat{q}) \ .$$

Proof. A Taylor-Lagrange expansion of order 1 can be derived for $h(\hat{s})$ since the derivative of h

$$h'(x) = -\frac{(2/3)n\overline{M}\vartheta x^2 + 4n\vartheta^2 x}{(2\vartheta^2 + (2/3)\overline{M}\vartheta x)^2}\exp\left(-\frac{nx^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta x}\right)\mathbb{1}_{x\geq 0} \ ,$$

is well defined for every $x \in \mathbb{R}$ (in particular, the left-side and right-ride derivatives at $x = 0$ coincide).
Therefore $h(\hat{s})$ equals

$$\begin{aligned}&h(s) + h^{'}(\tilde{s})(\hat{s} - s) \\ &\qquad = \exp\left(-\frac{ns^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta s}\right)\left[1 + \exp\left(g(s) - g(\tilde{s})\right)g'(\tilde{s})\mathbb{1}_{\tilde{s}\geq 0}(\hat{s} - s)\right] \ ,\end{aligned} \qquad \text{(A.34)}$$

where

$$s = L^2 - \frac{q}{n} - \frac{b_P^{(2)}}{n-1} \ , \quad \tilde{s} = L^2 - \frac{\tilde{q}}{n} - \frac{b_P^{(2)}}{n-1} \ , \quad \tilde{q} \in (q \wedge \hat{q}, q \vee \hat{q}) \ ,$$

and $s \geq 0$ because of the assumption $n > (q + m_P^{(2)})L^{-2}$.

For every $x, y > 0$, $|g'(x)| \leq 3n/(2\overline{M}\vartheta)$ and then $|g(x) - g(y)| \leq 3n|x - y|/(2\overline{M}\vartheta)$. It follows

$$|g'(\tilde{s})| \leq \frac{3n}{2\overline{M}\vartheta} \ , \qquad \text{(A.35)}$$

$$\exp\left(g(s) - g(\tilde{s})\right) \leq \exp\left(\frac{3n}{2\overline{M}\vartheta}|s - \tilde{s}|\right) = \exp\left(\frac{3|\tilde{q} - q|}{2\overline{M}\vartheta}\right) \quad .$$

Lemma A.6 is proved. □

□

References

[1] N. Aronszajn. Theory of reproducing kernels. May 1950.

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. 2013.

[3] S. Boucheron and M. Thomas. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 12, 2012.

[4] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious gaussian process models. 2012.

[5] J.A. Cuesta-Albertos, R. Fraiman, and T. Ransford. Random projections and goodness-of-fit test in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, pages 1–25, June 2006.

[6] A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schoelkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19 of *MIT Press, Cambridge*, pages 513–520, 2007.

[7] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, March 2012.

[8] A. Gretton, K. Fukumizu, Z. Harchaoui, and B.K. Sriperumbudur. A fast, consistent kernel two-sample test. 2009.

[9] N. Henze and B. Zirkler. A class of invariant and consistent tests for multivariate normality. *Comm. Statist. Theory Methods*, 19:3595–3617, 1990.

[10] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 1948.

[11] E.L. Lehmann and J. P. Romano. *Testing Statistical hypotheses*. Springer, 2005.

[12] K.V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519–530, 1970.

[13] S. Nikolov. Principal component analysis : Review and extensions. 2010.

[14] V.Y. Pan and Z.Q. Chen. The complexity of the matrix eigenproblem. *STOC '99 Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 507–516, 1999.

[15] B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. 1997.

[16] R. J. Serfling. *Approximation Theorems for Mathematical Statistics*. John Wiley & Sons, 1980.

[17] B.K. Sriperumbudur, A. Gretton, K. Fukumiju, B. Schölkopf, and G.R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, pages 1517–1561, 2010.

[18] M.S. Srivastava, S. Katayama, and Y. Kano. A two-sample test in high dimensional data. *Journal of Multivariate Analysis*, pages 349–358, 2013.

[19] G.J. Szekely and R.L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

[20] L. Zwald. *Performances d'Algorithmes Statistiques d'Apprentissage: "Kernel Projection Machine" et Analyse en Composantes Principales à Noyaux*. 2005.